NONSUPERVISED SEQUENTIAL CLASSIFICATION AND RECOGNITION OF PATTERNS

BY
E. A. PATRICK
AND
J. C. HANCOCK

Reprinted from IEEE TRANSACTIONS ON *INFORMATION THEORY*
Vol. IT-12, Number 3, July 1966
Pp. 362-372

Copyright 1966, and reprinted by permission of the copyright owner
PRINTED IN THE U.S.A.

Nonsupervised Sequential Classification and Recognition of Patterns

E. A. PATRICK, MEMBER, IEEE, AND J. C. HANCOCK, MEMBER, IEEE

Abstract—A Bayes approach to nonsupervised pattern recognition is given where n l-dimensional vector samples X_1, X_2,..., X_n are received unclassified, i.e., any one of M pattern sources ω_1, ω_2,..., ω_M, with corresponding probabilities of occurrence Q_{1_0}, Q_{2_0},..., Q_{M_0}, caused each sample X_s, $s = 1,2,...,n$.

The approach utilizes the fact that the cumulative distribution function (c.d.f.) of X_s is a mixture c.d.f., $F(X_s) = \sum_{i=1}^{M} F(X_s|\omega_i)\, Q_{i_0}$. It is assumed that available a priori knowledge includes knowledge of M and the family $\{F(X_s|\omega_i)\}$, where $F(X_s|\omega_i)$ is characterized by a vector B_{i_0}. In general, B_{i_0} and Q_{i_0}, $i = 1,2,...,M$ are considered fixed but unknown, and conditional probability of error in deciding which source caused X_n is minimized.

When the functional form of $F(X_s|\omega_i)$ in terms of B_{i_0} is unknown, the family $\{F(X_s|\omega_i)\}$ is taken to be the family of multinomial c.d.f.'s—an application of the histogram concept to the nonsupervisory problem. Additional nonparameteric a priori knowledge about the family—such as $F(X_s|\omega_i)$ is symmetrical, and/or $F(X_s|\omega_i)$ differs from $F(X_s|\omega_j)$ only by a translational vector—can be utilized in the Bayes solution.

I. Introduction

A. The Problem

LET THERE BE M possible pattern sources, ω_1, ω_2, $\cdots$ ω_M, only one of which is active, to produce an l-dimensional vector sample X_s at the receiver as shown in Fig. 1. It is assumed that if a pattern source ω_i is active on the sth sample, then the cumulative distribution function (c.d.f.) of X_s is $F(X_s \mid \omega_i)$. Let $P(\omega_i) = Q_{i_0}$ be the probability that the ith pattern source is active causing X_s, and assume Q_{i_0} fixed—independent of s.

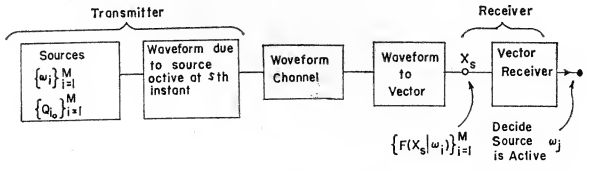

Fig. 1. Channel model.

We consider the problem where n vector samples, X_1, X_2, $\cdots$, X_n (let $Y_n = X_1, X_2, \cdots X_n$) are received unclassified (i.e., it is not known at the receiver which pattern source is active to produce sample X_s, $s = 1, 2, \cdots n$) and a decision is to be made on the nth sample, with minimum risk, as to which pattern source is active on this nth sample. We are especially interested in a general approach including the case when the family $\{F(X_s \mid \omega_i)\}$ is unknown and the a priori source probabilities $\{Q_{i_0}\}_{i=1}^{M}$ are unknown.

To help illustrate the problem further, consider the following example with two sources ($M = 2$) and one-dimensional vectors at the receiver ($l = 1$). When source ω_1 is active, assume the waveform at the channel input is a constant m_{1_0} and when source ω_2 is active, a constant m_{2_0}. Source ω_1 is assumed active with probability Q_{1_0} while ω_2 is assumed active with probability $(1 - Q_{1_0})$. Assume that the waveform channel simply consists of additive, Gaussian white noise with no memory such that the probability density function of X_s, given source ω_1 as active, is Gaussian with mean m_{i_0} and variance σ_0.

For this example, the channel model of Fig. 1 reduces to that shown in Fig. 2. Experimentally, given X_1, X_2, $\cdots$, X_n, one problem is to "learn" or estimate m_{1_0}, m_{2_0}, σ_0, and Q_{1_0}. Another problem is, given X_1, X_2, $\cdots$, X_n, to make a decision with minimum probability of error as to which source was active to cause sample X_n.

It will be advantageous to refer back to the example illustrated in Fig. 2 in a later section where computer simulated results are presented. Unless otherwise indicated, however, the model shown in Fig. 1 is under consideration.

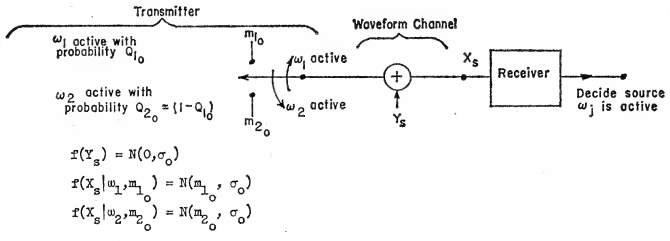

Fig. 2. Binary ($M = 2$) one-dimensional ($l = 1$) example.

B. Approach

We will call our approach a mixture [10] approach to nonsupervisory problems because the c.d.f. of X_s is

$$F(X_s) = \sum_{i=1}^{M} F(X_s \mid \omega_i) Q_{i_0}. \tag{1}$$

In order to apply classical Bayes results to the problem, we define a "parameter-conditional mixture,"

Manuscript received August 14, 1964; revised July 22, 1965, and January 27, 1966. The research reported herein is part of E. A. Patrick's work for the Ph.D. dissertation, School of Electrical Engineering, Purdue University, November 1965, and was supported under NASA Grant NsG 553.

The authors are with Purdue University, Lafayette, Ind.

$$F(X_s \mid B_0) = \sum_{i=1}^{M} F(X_s \mid \omega_i, B_{i0})Q_{i0} \qquad (2)$$

where B_{i_0} contains all the parameters characterizing $F(X_s \mid \omega_i)$ and B_0 contains all parameters in $B_{1_0}, B_{2_0}, \cdots, B_{M_0}$ and $Q_{1_0}, Q_{2_0}, \cdots, Q_{M_0}$.

Goal 1

For the case when $F(X_s \mid \omega_i)$ is unknown, we approximate it by a multinomial c.d.f., thus obtaining a set of parameters B_{i_0} which characterizes $F(X_s \mid \omega_i)$. This might be called an application of the histogram concept to nonsupervisory problems. We also give a construction technique for utilizing such nonparametric a priori knowledge as $F(X_s \mid \omega_i)$ is symmetrical or $F(X_s \mid \omega_i)$ differs from $F(X_s \mid \omega_j)$ $j \neq i$ only by a translation vector.

Goal 2

As defined, B_0 contains all the parameters characterizing this nonsupervisory problem, including $Q_{1_0}, Q_{2_0}, \cdots, Q_{M_0}$. As a second goal, we want to treat all parameters in B_0 as fixed but unknown, and show under what sufficient conditions *all* these parameters can be "learned" as well as give the minimum conditional probability of error solution. Consistent with a Bayes approach, we will let B, B_i, and Q_i, be random variables corresponding to our uncertainty in B_0, B_{i_0}, and Q_{i_0}, respectively.

Discussion of Goal 2

Spragins [17] considered a convergence theorem giving sufficient conditions (i and ii below) for a Bayes solution to converge.

i) There exists a sequence of functions $f_n(Y_n)$ converging to B_0 with probability one.

ii)
1) $f(B) > 0$ in some sphere containing B_0.
2) The a posteriori density $f(B \mid Y_n)$ is calculated by the Bayes rule.

The solution to the nonsupervisory problem presented in this paper is a Bayes solution; consequently ii)2 above is satisfied. We assume ii)1 is satisfied. The approach taken in this paper makes it possible to examine sufficient conditions for i) to be satisfied. These sufficient conditions are specified in terms of the a priori knowledge defining the nonsupervisory problem.

a) The number of pattern sources M
b) The family $\{F(X_s \mid \omega_i)\}$
c) Any constraints on B or X_s sufficient for identifiability (defined in Appendix I)
d) Any constraints additional to c) required in order to exhibit a consistent estimator for B_0, thus satisfying i) above.

In summary, the nonsupervisory problem is partly defined in terms of a priori knowledge a) and b). In terms of this a priori knowledge, sufficient constraints (if they exist) are determined to obtain property c)—identifiability. Then, further constraints are imposed to obtain d)—i.e., to insure that a consistent estimator for B_0 can be found. After all this, we are assured that given a priori knowledge a), b), c), and d), the Bayes solution will converge.

In this paper, we will allude to the references concerning d), a research problem where there are only a few results. In Appendix I, we consider sufficient constraints c) for identifiability for several nonsupervisory problems.

C. Literature Review—Bayes Approach to Nonsupervised Pattern Recognition

An optimality criterion frequently used is as follows. Given Y_{n-1}, make a decision as to which of M pattern sources caused sample X_n. This decision is made by a decision function obtained with a system constraint of minimum sample-conditional probability of error.[1] The word sample is used here to make clear that we are talking about probability of error conditioned on the past samples, Y_{n-1}.

Abramson and Braveman [1] considered a problem where it is known which pattern source caused samples X_s, $s = 1, 2, \cdots, n_i$ (i.e., the samples are classified). That is, the a priori knowledge includes knowledge that

$$F(X_s \mid B_0) = F(X_s \mid \omega_i, B_{i_0}),\ i \text{ known},\ s = 1, 2 \cdots, n_i.$$

Further, it is known that the family $\{F(X_s \mid \omega_i)\}$ is a multidimensional Gaussian family, with only the mean vector m_{i_0} (in B_{i_0}) unknown for each member of the family. The authors also assume a c.d.f. $F(B)$ is available describing the a priori uncertainty in the unknown parameters. Using all this a priori knowledge, they obtained a system minimizing sample-conditional probability of error.

Keehn [15] extended the work of Abramson and Braveman to the case where the mean vector and covariance matrix (in B_{i_0}) are unknown. He carefully defined c.d.f.'s $F(B_i)$ for all i such that the a posteriori c.d.f. of B_i, for each i is reproducing [20].

Daly [3] investigated a nonsupervisory system where the classification of the samples is unknown. A priori knowledge includes: knowledge that there are M pattern sources, the family $\{F(X_s \mid \omega_i)\}$ is known, the set of mixing parameters $\{Q_{i_0}\}_{i=1}^{M}$ are known, and a c.d.f. $F(B)$ is available. Daly computed the decision function which minimizes the sample-conditional probability of error for decision on sample X_n. Application of this decision function requires computation of $f(X_n, \omega_i \mid Y_{n-1})$, $i = 1, 2, \cdots M$. His computation of $f(X_n, \omega_i \mid Y_{n-1})$ is a sum of $M^{(n-1)}$ terms, thus requiring rapidly increasing computer memory. He did not consider sufficient constraints on the parameters for the system to converge.

Fralick, [2], [14] looking for an iterative solution to Daly's problem, obtained an iterative form for $f(B_i \mid Y_{n-1})$ assuming that if B_{i_0} characterizes $F(X_s \mid \omega_i)$ and B_{j_0}

[1] More generally, sample conditional risk can be minimized. We will, for practical reasons, be concerned with sample-conditional probability of error.

characterizes $F(X_s \mid \omega_i)$, then $F(B_i \mid Y_{n-1}, B_i) = F(B_i \mid Y_{n-1})$.

Patrick and Hancock [18], [16] showed more generally, without making the assumption $F(B_i \mid Y_{n-1}, B_i) = F(B_i \mid Y_{n-1})$, that the desired a posteriori probability density $f(B_i \mid Y_{n-1})$ is either of the growing form or, equivalently, is computed by integrating the joint density $f(B \mid Y_{n-1})$ with respect to all vectors except B_i, where $f(B \mid Y_{n-1})$ has an iterative form.

Some of the first work on applying a histogram, approximating $F(X_s \mid \omega_i)$, to adaptive communication systems was done by Sebestyen [5], [8]. He considered only classified samples. Patrick and Hancock [6], [7] introduced the concept of a histogram, approximating $F(X_s \mid \omega_i)$, to the nonsupervisory problem. They presented computer simulated results [7] showing rates of convergence for a few examples.

Teicher [9], [10] defined a mixture c.d.f. and identifiability, and gave a theorem giving sufficient conditions for a mixture to be identifiable. Identifiability assures that a unique solution B_0 of $F(X_s) = F(X_s \mid B)$ exists, where $F(X_s)$ is a mixture, and with a few additional constraints d), frequently permits exhibition of a consistent estimator for B_0. In Appendix I we include and give a simple extension of some of Teicher's work, and define a parameter conditional mixture $F(X_s \mid B)$ which is a useful concept for applying Bayes Theorem to mixtures. In addition, we state several propositions giving sufficient conditions for a mixture to be identifiable.

Among others who have been concerned with identifiability c) and consistent estimators d) are Cooper and Cooper [4] and Robbins [12].

Scudder [19], [20], Jakowatz, Shuey, and White [21], and Proakis and Drouilhet [23] have investigated "Decision Directed Measurement" techniques when samples are unclassified. It would be of value to indicate the a priori knowledge assumed in the design of such systems and to list this a priori knowledge a), b), c), and d) as in Section *I. B.* It may be that such systems are optimum against the a priori knowledge they utilize, but do not use all available a priori knowledge. A system not using certain a priori knowledge may be simplified and better in some overall cost consideration even though it does not minimize conditional probability of error against all available a priori knowledge.

Among others who have contributed to adaptive communication systems, Kailath [22] gave an estimator-correlator interpretation to a minimum probability of error system and extended the interpretation to adaptive systems.

II. Mixtures and a Construction Technique

A. Mixtures and Parameter Conditional Mixtures

In this section we show that the c.d.f. of X_s, when X_s is unclassified, is a mixture c.d.f. We then define a parameter-conditional mixture in anticipation of the Bayes solution in Section III.

A mixture results when a vector X_s can be partitioned M ways, $\omega_1, \omega_2, \cdots \omega_M$, as illustrated in Fig. 1. For example, define by (X_s, ω_i) the event that X_s is caused by pattern source ω_i being active. Since only one of the M pattern sources can be active to cause X_s, there are M possible mutually exclusive and exhaustive events, $(X_s, \omega_1), (X_s, \omega_2), \cdots, (X_s, \omega_M)$. If $P(\omega_i) = Q_i$, independent of s,

$$F(X_s) = \sum_{i=1}^{M} F(X_s \mid \omega_i) Q_i. \tag{3}$$

When we speak of a family of Gaussian c.d.f.'s or a family of multinomial c.d.f.'s, we have in mind the nature of the parameters which characterize the family. It is therefore appropriate to define a parameter-conditional mixture c.d.f. $F(X_s \mid B)$ constructed using the collection $\{F(X_s \mid \omega_i, B_i\}$ where $F(X_s \mid \omega_i)$ is characterized by B_i. To do this, define

$$B = B_1 \cup B_2 \cup \cdots \cup B_M \cup B_{M+1} \tag{4}$$

where

B_i: vector characterizing $F(X_s \mid \omega_i)$
$B_i \cup B_j$: collection of all entries in both B_i and B_j
$B_{M+1} = \{Q_i\}_1^M$. (5)

Thus, B is simply the collection of the mixing parameters and all entries in $B_1, B_2, \cdots B_M$. In other words, B contains all fixed but unknown parameters characterizing the problem.

Since $(X_s, \omega_1), (X_s, \omega_2), \cdots, (X_s, \omega_M)$ are mutually exclusive and exhaustive events,

$$F(X_s \mid B) = \sum_{i=1}^{M} F(X_s \mid \omega_i, B) P(\omega_i \mid B). \tag{6}$$

Now, $F(X_s \mid \omega_i)$ is completely characterized by B_i, therefore

$$F(X_s \mid \omega_i, B) = F(X_s \mid \omega_i, B_i) \tag{7}$$

and since B contains Q_i,

$$P(\omega_i \mid B) = P(\omega_i) = Q_i. \tag{8}$$

Thus, (5) becomes[2]

$$F(X_s \mid B) = \sum_{i=1}^{M} F(X_s \mid \omega_i, B_i) Q_i. \tag{9}$$

Given $F(X_s)$, when does $F(X_s) = F(X_s \mid B)$ have a unique solution B_0? The answer is that there is a unique solution B_0 when the class of mixtures is identifiable, several sufficient conditions for which are given in Appendix I. Given identifiability c), a consistent minimum distance estimator [18] for B_0 can be exhibited when $F(X_s \mid \omega_i, B_i)$ is continuous in X_s and B_i—thus establishing property d). As stated earlier, with c) and d) satisfied, the Bayes solution will converge.

[2] For known parameters, (9) becomes

$$F(X_s \mid B_0) = \textstyle\sum_{i=1}^{M} F(X_s \mid \omega_i, B_{i_0}) Q_{i_0}.$$

In the next section, we give sufficient constraints for c) and d) to be satisfied when $F(X_s \mid \omega_i)$ is approximated by a multinomial c.d.f.

B. The Fixed Bin Model

A priori knowledge of the family $\{F(X_s \mid \omega_i)\}$ is required before constructing the parameter-conditional mixture (9). Our purpose now is to apply the histogram concept to this nonsupervisory problem, for situations where the family $\{F(X_s \mid \omega_i)\}$ is unknown. To do this, we develop a construction method where a multinomial c.d.f. approximates $F(X_s \mid \omega_i)$, utilizing nonparametric a priori knowledge about $F(X_s \mid \omega_i)$ for each i. Thus we will have achieved the first goal of this paper.

When samples $X_1, X_2, \cdots X_n$ are all from the same class, say ω_i, it is well known that a histogram can be formed from these samples to approximate $f(X_s \mid \omega_i)$. In a nonsupervisory problem where the samples are unclassified, a certain sufficient amount of a priori knowledge is required if a histogram for $f(X_s \mid \omega_i)$ is to be obtained. We will now give a construction method which leads to a determination of such a sufficient amount of a priori knowledge.

Being more general than in the previous Section, let X_s be a sequence of v l-dimensional samples, $X_{s_1}, X_{s_2}, \cdots X_{s_v}$, with the same pattern class active for all v samples.

Assume further that, given ω_i and B_i, the v samples are statistically independent. In order to keep the same notation as in A, let $X_s = \{X_{s_k}\}_1^v$, and let $Y_n = X_1, X_2, \cdots X_n$.

X_{s_k} is an l-dimensional vector. We quantize each of these dimensions into R levels, obtaining R^l l-dimensional "cubes" or "bins". Each l-dimensional bin is assumed to have the same volume. X_{s_k} can lie in any of these R^l bins, or in the $(R^l + 1)$st bin representing the remaining part of the l-dimensional space. The bins are indexed and indicated by I_ξ, $\xi = 1, 2, \cdots, (R^l + 1)$.

$F(X_s \mid \omega_i)$ is now approximated using a multinomial c.d.f. characterized by the set of parameters $G^i = (p_1^i, p_2^i, \cdots, p_{R^l}^i)$ where p_ξ^i is the probability X_{s_k} falls in bin I_ξ, given pattern class ω_i is active. The probability that X_{s_k} falls in bin I_{R^l+1}—given pattern class ω_i is active (denoted by $p_{(R^l+1)}^i$) is given, in terms of $p_1^i, \cdots, p_{R^l}^i$ by

$$p_{(R^l+1)}^i = 1 - \sum_{\xi=1}^{R^l} p_\xi^i, \qquad i = 1, 2, \cdots M. \tag{10}$$

Since X_s is a sequence of v samples, v of the $(R^l + 1)$ bins receive samples during the sth sequence, not all bins being necessarily different. Let this relative frequency in the bins during the sth sequence be denoted by

$$V_s = (v_{s_1}, v_{s_2}, \cdots, v_{s_{(R^l+1)}}), \qquad v_{s_\xi} = 0, 1, \cdots, v. \tag{11}$$

The probability distribution of V_s, given pattern class ω_i and G^i, is multinomial:

$$P(V_s \mid \omega_i, G^i) = \frac{v!}{v_{s_1}! \cdots v_{s_{(R^l+1)}}!} \prod_{\xi=1}^{R^l+1} [p_\xi^i]^{v_{s_\xi}} \tag{12}$$

and analogous to (8),

$$P(V_s \mid B) = \sum_{i=1}^{M} P(V_s \mid \omega_i, B_i) Q_i \tag{13}$$

where

$$\begin{aligned} B_i &= G^i \\ B &= (G^1, G^2, \cdots G^M, Q_1, \cdots, Q_M). \end{aligned} \tag{14}$$

An example of the fixed bin model for $l = 1$, $M = 2$, and $v = 1$ is shown in Fig. 3. For this case, (12) and (13) combine to give

$$p_\xi^0 = \sum_{i=1}^{M} p_\xi^i Q_i, \qquad \xi = 1, 2, \cdots, R^l + 1$$

where

$$p_\xi^0 = P[X_{s_k} \text{ falls in Bin } I_\xi].$$

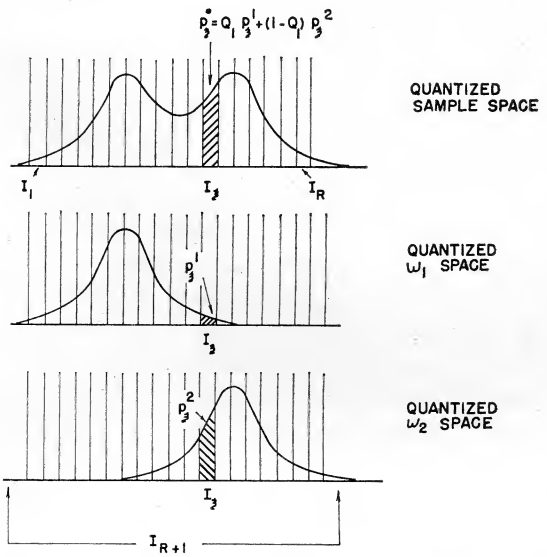

Fig. 3. Quantized spaces, $l = 1$, $v = 1$, $M = 2$.

C. Sufficient a priori Knowledge for Identifiability

As stated in Section II–A, a nonsupervisory problem will be said to be identifiable if, given $F(X_s)$ and the parameter conditional mixture $F(X_s \mid B)$,

$$F(X_s) = F(X_s \mid B) \tag{15}$$

has a unique solution B_0.

It is possible to consider particular examples of nonsupervisory problems and decide if they are identifiable. We will define a particular nonsupervisory problem according to the a priori knowledge required to construct $F(X_s \mid B)$ (9) and according to constraints on the parameter space B and/or the sample space X_s. That is, a particular nonsupervisory problem is defined by

a) the number of pattern classes M
b) the family $\{F(X_s \mid \omega_i)\}$
c) any constraints on B or X_s sufficient for identifiability
d) any additional constraints to insure the existence of a consistent estimator for B_0.

If, using a) and b) to construct $F(X_s \mid B)$ and imposing the constraints c), (15) has a unique solution B_0, then it will be said that sufficient a priori knowledge exists for identifiability.

For example, if $f(X_s \mid \omega_i, B_{i_0})$ is one-dimensional Gaussian with $B_{i_0} = (m_{i_0}, \sigma_{i_0})$ and $M = 2$, Proposition 1 shows that for identifiability it is sufficient that $\sigma_{1_0} \neq \sigma_{2_0}$, or if $\sigma_{1_0} = \sigma_{2_0}$, $m_{1_0} \neq m_{2_0}$ be constraints on B_0.

Further, for the binary ($M = 2$) on-off case ($m_{2_0} = 0$ and this is known), no constraints on B_0 are required.

For the case where the family $\{F(X_s \mid \omega_i)\}$ is multinomial or where the fixed bin model is used because the family is unknown, we are concerned with the parameter conditional mixture (13). Since X_s is a discrete vector, identifiability requires that

$$P(X_s) = P(X_s \mid B)$$

have a unique solution B_0 where $P(X_s \mid B)$ is a parameter conditional mixture of multinomial distributions. Proposition 5 shows that, for identifiability, it is sufficient that $v \geq 2M - 1$ where v is such that

$$X_s = \{X_{sk}\}_{k=1}^{v}.$$

This is the reason why, when introducing the fixed bin model, we made provision for taking v samples in the sth sequence with the same pattern class active. In particular, for the binary case ($M = 2$) it is sufficient to take 3 or more samples while the same pattern class is active.

In general, the number of parameters in B_i is quite large when the fixed bin model is used. This is because the fixed bin model assumes little knowledge about the family $\{F(X_s \mid \omega_i)\}$—only that $F(X_s \mid \omega_i)$ can be approximated by a multinomial c.d.f. On the other hand, if the family $\{F(X_s \mid \omega_i)\}$ is known a priori to be Gaussian, B_i contains a mean vector and a covariance matrix—usually fewer parameters than for the fixed bin model case. This would seem to suggest that the less known a priori, the more parameters will be required to characterize the problem.

The fixed bin model introduces a construction technique for utilizing a priori knowledge. In the following section, we will show how such a priori knowledge as $F(X_s \mid \omega_i)$ is symmetrical, and/or $F(X_s \mid \omega_i)$ differs from $F(X_s \mid \omega_j)$, $j \neq i$, only by a translational vector, can be taken into account. Utilizing this additional a priori knowledge reduces the number of fixed but unknown parameters in B. We will also show how to take into account a priori knowledge that $f(X_s \mid \omega_i) = 0$ for X_s in a given region of the sample space. This will be done for the binary case ($M = 2$).

D. Utilizing Additional a priori Knowledge About $F(X_s \mid \omega_i)$

If it is known, for example, that $F(X_s \mid \omega_i)$ is symmetrical, then approximating this c.d.f. by a multinomial c.d.f. does not utilize the symmetrical knowledge. For this case, we would use an appropriately defined symmetrical multinomial distribution to approximate the c.d.f. If, as another example, it is known that $F(X_s \mid \omega_i)$, $i = 1, 2, \cdots, M$, differs only by translational vectors, we would approximate each $F(X_s \mid \omega_i)$ by an appropriately defined translated multinomial c.d.f.

We have not yet said how we propose to count the bins in l-dimensional space, although writing $p_1^i, p_2^i, \cdots p_{R^l}^i$ indicates we must have had some counting procedure in mind. One method of counting is to redefine p_ξ^i

$$p_\xi^i, = p_{i_1, i_2, \cdots, i_l}^i \qquad 1 \leq j_a \leq R \text{ for all } j_a. \tag{16}$$

Then let

$$\begin{aligned} p_1^i &= p_{1,1,\cdots,1}^i \\ p_R^i &= p_{R,1,\cdots,1}^i \\ p_{R+1}^i &= p_{R,2,1,\cdots,1}^i. \end{aligned} \tag{17}$$

Denote the family $\{F(X_s \mid \omega_i, B_i)\}$ where $B_i = G^i$ by $\mathfrak{F}_P = \{F(X_s \mid \omega_i, G^i)\}$. This is the family used in the construction of the parameter-conditional mixture (13). Next, define the family $\mathfrak{F}_{TP}$ of multinomial c.d.f.' differing only by translational vectors, $\{\theta_i\}$. To accomplish this, define G_{θ_0} where θ_0 is a vector of l indexes.

$$\theta_0 = \left(\frac{R+1}{2}, \frac{R+1}{2}, \cdots, \frac{R+1}{2},\right) R \text{ odd}.$$

The vector θ_0 locates the center bin in the l-dimensional space with R^l quantum levels used for representing G_{θ_0}. In terms of G_{θ_0}, the vector G^i characterizing $F(X_s \mid \omega_i)$ is expressed as

$$G^i = G_{(\theta_0 - \theta_i)}, \quad \text{with} \quad p_{R^l+1}^i = 0, \qquad i = 1, 2, \cdots, M. \tag{18}$$

Also define a family $\mathfrak{F}_{STP}$ of symmetrical multinomial c.d.f.'s differing only by translational vectors, $\{\theta_i\}$ by letting $G_{S\theta_0}$ be a vector whose entries are symmetrical about θ_0 in each of the l dimensions.

Now, as an example, assume that it is known that $F(X_s \mid \omega_i)$, $i = 1, 2, \cdots M$ are all identical except for different translational vectors. We approximate these c.d.f.'s my members of the family $\mathfrak{F}_{TP}$.

Accordingly, X_s is quantized, and the relative frequency vector V_s is thus obtained. The probability density of V_s, given B, is (13).

$$P(V_s \mid B) = \sum_{i=1}^{M} P(V_s \mid \omega_i, B_i) Q_i \tag{19}$$

where

$$\begin{aligned} B_i &= (G_{\theta_0}, \theta_i) \\ B &= (G_{\theta_0}, \theta_1, \cdots, \theta_M, Q_1, \cdots Q_M). \end{aligned} \tag{20}$$

Comparing (20) with (14) we find that the a priori knowledge that $\{F(X_s \mid \omega_i)\} \in \mathfrak{F}_{TP}$ reduced the number of parameters in B by $(M-1)R^l - M$ as compared to when $\{F(X_s \mid \omega_i)\} \in \mathfrak{F}_P$. If $\{F(X_s \mid \omega_i)\} \in \mathfrak{F}_{STP}$ instead of $\mathfrak{F}_{TP}$, the number of parameters characterizing the mixture c.d.f. is further reduced by $[(R+1)/2 - 1]^l$, for R odd.

E. Family of Multinomial C.D.F.'s with Spacial Constraints and $v = 1$

Let $l = 1$, $M = 2$, $v = 1$, and assume it known a priori that $F(X_s \mid \omega_2) = 0$ for $X_s \leq \theta_1$ and that $F(X_s \mid \omega_1) = 1$ for $X_s \geq \theta_2$, where θ_1 and θ_2 are translational parameters. This "spacial constraint" corresponds to an approximation that can be made when the signal-to-noise ratio is "sufficiently large," and $F(X_s \mid \omega_i)$ is symmetrical about its translational parameter θ_i.

Using the fixed-bin model, let $\{F(X_s \mid \omega_i)\} \in \mathfrak{F}_{SP}$, the family of symmetrical multinomial c.d.f.'s. Since $v = 1$, all entries in V_s will be zero except one entry which will be unity. If $X_s \leq \theta_1$, it is classified as from pattern class ω_1; if $X_s \geq \theta_2$, it is classified as from pattern class ω_2, but if $\theta_1 < X_s < \theta_2$, it is unclassified. Thus, the probability distribution of V_s, given B, is

$$P(V_s \mid B) = \begin{cases} P(V_s \mid \omega_1, B_1), & X_s \leq \theta_1 \\ P(V_s \mid \omega_2, B_2), & X_s \geq \theta_2 \\ \sum_{i=1}^{2} P(V_s \mid \omega_i, B_i)Q_i, & \theta_1 < X_s < \theta_2 \end{cases} \tag{21}$$

where

$$\begin{aligned} B_i &= (G^i_{S,\theta_i}, \theta_i), \qquad i = 1, 2 \\ B &= (G^1_{S,\theta_1}, G^2_{S,\theta_2}, \theta_1, \theta_2, Q_1) \end{aligned} \tag{22}$$

and G^i_{S,θ_i} is a vector of probabilities $p^i_1, \cdots, p^i_R$, R odd, symmetric about the middle entry $p^i_{(R+1)/2}$.

An example using the above a priori knowledge where samples greater than θ_1 but less than θ_2 were not used is given in Patrick and Hancock [7]. Histograms for $F(X_s \mid \omega_1)$ and $F(X_s \mid \omega_2)$ were obtained while using sample quantiles as estimators for θ_{1_0} and θ_{2_0}. In this example, $Q_{1_0} = \frac{1}{2}$ and was known a priori, but θ_{1_0}, θ_{2_0}, $F(X_s \mid \omega_1)$ and $F(X_s \mid \omega_2)$ were unknown a priori. Average error vs. n was obtained by computer simulation with $F(X_s \mid \omega_i)$ Gaussian (but unknown at the receiver). This average error is shown to converge, for practical purposes, to the probability if error obtained when θ_{1_0}, θ_{2_0} are known, and $F(X_s \mid \omega_i)$ known Gaussian. Equation (21), however, when used in the minimum sample-conditional probability of error solution presented next in this paper, shows how to use those samples greater than θ_1 and less than θ_2, and does not require external estimators for θ_{1_0} and θ_{2_0}.

III. Minimum Conditional Probability of Error

A. Introduction

We are interested in observing sample X_n and deciding which pattern source ω_i caused X_n. For each pattern source ω_i and a given decision function $d(X_n)$, we define the conditional loss $L(d(X_n) \mid \omega_i)$, independent of B and $Y_{n-1} = \{X_s\}_1^{n-1}$. Also, we define the conditional risk $r(d \mid \omega_i)$ as the conditional loss averaged over all values of X_n.

It is shown in Patrick [18] that if $f(B \mid Y_{n-1})$ is the sample-conditional density of B, then the sample conditional risk is

$$r(d \mid Y_{n-1}) = \int dB \left\{ \int dX_n \cdot \left[\sum_{i=1}^{M} L(d(X_n) \mid \omega_i) f(X_n \mid \omega_i, B_i) Q_i \right] \right\} f(B \mid Y_{n-1}). \tag{23}$$

It is well known that for a 0, 1 loss function, the decision function minimizing (23) is

$$d(X_n) : \text{choose } \omega_i \text{ such that}$$
$$f(X_n, \omega_i \mid Y_{n-1}) = \operatorname*{Sup}_i \{f(X_n, \omega_i \mid Y_{n-1})\}. \tag{24}$$

For this loss function, minimum conditional risk is equivalent to minimum conditional probability of error.

When X_n is a discrete random vector, the decision function for 0, 1 loss function is

$$d(X_n) : \text{choose } \omega_i \text{ such that}$$
$$P(X_n, \omega_i \mid Y_{n-1}) = \operatorname*{Sup}_i \{P(X_n, \omega_i \mid Y_{n-1})\}. \tag{25}$$

B. Computation of $f(B \mid Y_{n-1})$ for Mixtures

In order to minimize sample-conditional probability of error, $f(B \mid Y_{n-1})$ must be computed using a priori knowledge a), b), c), and d) of Section II—C and $F(B)$—at least an appropriately defined uniform c.d.f., not ruling out the true value of B.

Working with density functions rather than c.d.f.'s, $f(B \mid Y_{n-1})$ is given by Bayes Theorem as follows:

$$f(B \mid Y_{n-1}) = \frac{f(X_{n-1} \mid B)(B \mid Y_{n-2})}{f(X_{n-1} \mid Y_{n-2})}. \tag{26}$$

The denominator on the right side of (26) is a normalization constant which assures that $f(B \mid Y_{n-1})$ integrates over the B space to unity. $f(B \mid Y_{n-2})$ is the density in the B space at the $(n-2)$ stage. $f(X_{n-1} \mid B)$ is a function directly utilizing the a priori knowledge above. It can be shown that the identifiability requirement c) is sufficient to guarantee the existence of an estimator for B_0, converging to B_0 with probability one, under rather general conditions. For example, Patrick [18] has exhibited a minimum distance estimator $\xrightarrow{P} B_0$ when $F(X_s \mid \omega_i, B_i)$ is continuous in X_s and B_i. Blischke [11] has exhibited consistent moment estimators for the parameters of a mixture of two ($M = 2$) binomial c.d.f.'s. His result immediately extends to a mixture of two ($M = 2$) multinomial c.d.f.'s.

Further research is required to determine more general conditions under which identifiability assures the existence of an estimator converging to B_0 with probability one. When such an estimator does exist, we are assured that $f(B \mid Y_{n-1})$ converges to a Dirac delta function [17] at B_0.

Using a priori knowledge a) and b) to construct $f(X_{n-1} \mid B)$ according to (9), (26) becomes

$$f(B \mid Y_{n-1}) = \frac{\left[\sum_{i=1}^{M} f(X_{n-1} \mid \omega_i, B_i)Q_i\right] f(B \mid Y_{n-2})}{f(X_{n-1} \mid Y_{n-2})}. \quad (27)$$

Equation (27) is a fundamental result for the a posteriori probability density of the vector B characterizing the parameter-conditional mixture. Note that B contains all parameters characterizing this nonsupervisory problem, including the mixing parameters $\{Q_i\}_{i=1}^{M}$.

Using the iterative solution for $f(B \mid Y_{n-1})$ in (27), the decision equation (24) is implemented by computing

$$f(X_n, \omega_i \mid Y_{n-1}) = \int Q_i f(X_n \mid \omega_i, B_i) f(B \mid Y_{n-1})\, dB \quad (28)$$

and the decision equation (25) is implemented by computing

$$P(X_n, \omega_i \mid Y_{n-1}) = \int Q_i P(X_n \mid \omega_i, B_i) f(B \mid Y_{n-1})\, dB. \quad (29)$$

C. Systems Minimizing Sample Conditional Probability of Error

In this section we consider the design of systems minimizing sample-conditional probability of error. When $F(X_s \mid \omega_i)$ is approximated by a multinomial c.d.f. using the fixed bin model, X_s is a discrete random vector. We therefore use decision equation (25) with $P(X_n, \omega_i \mid Y_{n-1})$ computed in terms of $f(B \mid Y_{n-1})$ by (29).

Using the fixed bin model, denote the bins that the v samples on the nth observation fall in by I_{η_k}, $k = 1, 2, \cdots, v$. Using this notation, we obtain from (12) and (29) that

$$P(X_n, \omega_i \mid Y_{n-1}) = v! \int \left[\left\{\prod_{k=1}^{v} p_{\eta_k}^i\right\} Q_i\right] f(B \mid Y_{n-1})\, dB. \quad (30)$$

It is convenient to define the sample-conditional expectation of

$$v!\left[\left\{\prod_{k=1}^{v} p_{\eta_k}^i\right\} Q_i\right] \quad \text{by} \quad a_{n-1}^i; \text{ that is}$$
$$a_{n-1}^i = E\left[\left\{v! \prod_{k=1}^{v} p_{\eta_k}^i\right\} Q_i \mid Y_{n-1}\right]. \quad (31)$$

If $v = 1$, (31) reduces to

$$a_{n-1}^i = E[p_\eta^i Q_i \mid Y_{n-1}] \quad (32)$$

where p_η^i is the probability X_n falls in bin I_η given that ω_i caused X_n.

Equation (31) used along with decision equation (25) has an interesting interpretation: To minimize sample-conditional probability of error while making a decision on the nth sample, observe the relative frequency vector V_n. Then compute the probability of this relative frequency vector given the past samples Y_{n-1} and ω_i, for each i, and make decisions as follows:

$$\text{choose } \omega_i \quad a_{n-1}^i = \operatorname{Sup}_i \{a_{n-1}^i\}. \quad (33)$$

When $F(X_s \mid \omega_i, B_i)$ is continuous in X_s and B_i, the decision equation is (24). For example, if the family is multinomial Gaussian,

$$f(X_{n-1} \mid \omega_i, B_i) = \frac{1}{((2\pi)^{l/2} |\Phi_{xx}^i|^{1/2})} \cdot \exp\{-\tfrac{1}{2}(X_{n-1} - \theta_i)^T [\Phi_{xx}^i]^{-1} (X_{n-1} - \theta_i)\}$$

where

$$\Phi_{xx}^i = E[X_s^T X_s \mid \omega_i] \quad \text{and} \quad \theta_i = E[X_s \mid \omega_i].$$

Note that

$$B_i = (\Phi_{xx}^i, \theta_i)$$
$$B = (\{\Phi_{xx}^i\}_{i=1}^{M}, \{\theta_i\}_{i=1}^{M}, \{Q_i\}_{i=1}^{M}).$$

The two types of optimum systems which make decisions with minimum sample-conditional probability of error are shown in Fig. 4. The upper system uses the fixed bin model, assuming the family $\{F(X_s \mid \omega_i)\}$ is multinomial and the lower system is for cases where the family is known and whose ith member is continuous in B_i and X_s.

Through (27)–(29), we have the minimum sample-conditional probability of error solution for this nonsupervisory problem and have thus obtained the second goal of the paper.

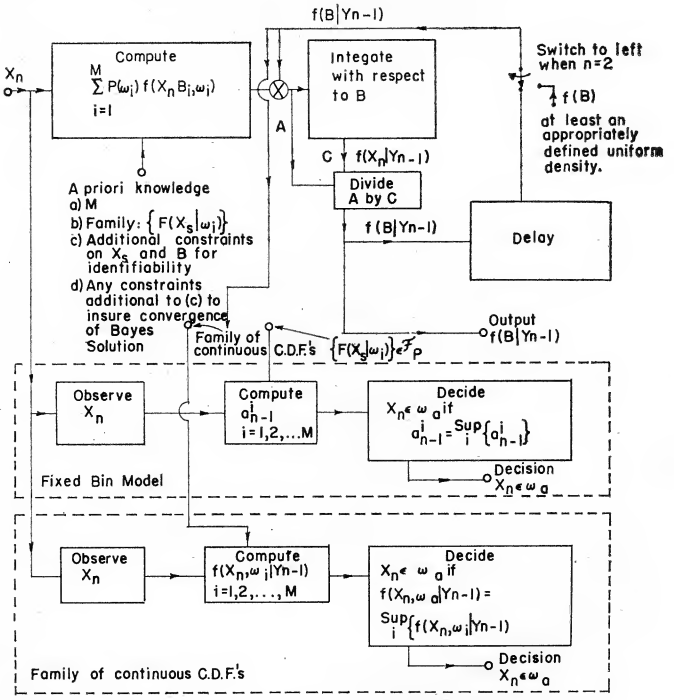

Fig. 4. Minimum conditional probability of error systems.

D. Marginal a posteriori Density of a Parameter in B

Sometimes it is desirable to obtain the a posteriori probability of just one parameter in B. The Bayes estimator of one parameter, for example, is calculated from the marginal a posteriori probability density of that parameter. Therefore, let γ_{k_i} be some parameter in B_k. The sample-conditional density of γ_{k_i} is obtained by

E. Family of Multinomial C.D.F.'s with Spacial Constraints and $v = 1$

Let $l = 1$, $M = 2$, $v = 1$, and assume it known a priori that $F(X_s \mid \omega_2) = 0$ for $X_s \leq \theta_1$ and that $F(X_s \mid \omega_1) = 1$ for $X_s \geq \theta_2$, where θ_1 and θ_2 are translational parameters. This "spacial constraint" corresponds to an approximation that can be made when the signal-to-noise ratio is "sufficiently large," and $F(X_s \mid \omega_i)$ is symmetrical about its translational parameter θ_i.

Using the fixed-bin model, let $\{F(X_s \mid \omega_i)\} \in \mathfrak{F}_{SP}$, the family of symmetrical multinomial c.d.f.'s. Since $v = 1$, all entries in V_s will be zero except one entry which will be unity. If $X_s \leq \theta_1$, it is classified as from pattern class ω_1; if $X_s \geq \theta_2$, it is classified as from pattern class ω_2, but if $\theta_1 < X_s < \theta_2$, it is unclassified. Thus, the probability distribution of V_s, given B, is

$$P(V_s \mid B) = \begin{cases} P(V_s \mid \omega_1, B_1), & X_s \leq \theta_1 \\ P(V_s \mid \omega_2, B_2), & X_s \geq \theta_2 \\ \sum_{i=1}^{2} P(V_s \mid \omega_i, B_i)Q_i, & \theta_1 < X_s < \theta_2 \end{cases} \tag{21}$$

where

$$\begin{aligned} B_i &= (G^i_{S,\theta_i}, \theta_i), \qquad i = 1, 2 \\ B &= (G^1_{S,\theta_1}, G^2_{S,\theta_2}, \theta_1, \theta_2, Q_1) \end{aligned} \tag{22}$$

and G^i_{S,θ_i} is a vector of probabilities $p^i_1, \cdots, p^i_R$, R odd, symmetric about the middle entry $p^i_{(R+1)/2}$.

An example using the above a priori knowledge where samples greater than θ_1 but less than θ_2 were not used is given in Patrick and Hancock [7]. Histograms for $F(X_s \mid \omega_1)$ and $F(X_s \mid \omega_2)$ were obtained while using sample quantiles as estimators for θ_{1_0} and θ_{2_0}. In this example, $Q_{1_0} = \frac{1}{2}$ and was known a priori, but θ_{1_0}, θ_{2_0}, $F(X_s \mid \omega_1)$ and $F(X_s \mid \omega_2)$ were unknown a priori. Average error vs. n was obtained by computer simulation with $F(X_s \mid \omega_i)$ Gaussian (but unknown at the receiver). This average error is shown to converge, for practical purposes, to the probability if error obtained when θ_{1_0}, θ_{2_0} are known, and $F(X_s \mid \omega_i)$ known Gaussian. Equation (21), however, when used in the minimum sample-conditional probability of error solution presented next in this paper, shows how to use those samples greater than θ_1 and less than θ_2, and does not require external estimators for θ_{1_0} and θ_{2_0}.

III. Minimum Conditional Probability of Error

A. Introduction

We are interested in observing sample X_n and deciding which pattern source ω_i caused X_n. For each pattern source ω_i and a given decision function $d(X_n)$, we define the conditional loss $L(d(X_n) \mid \omega_i)$, independent of B and $Y_{n-1} = \{X_s\}_1^{n-1}$. Also, we define the conditional risk $r(d \mid \omega_i)$ as the conditional loss averaged over all values of X_n.

It is shown in Patrick [18] that if $f(B \mid Y_{n-1})$ is the sample-conditional density of B, then the sample conditional risk is

$$r(d \mid Y_{n-1}) = \int dB \left\{ \int dX_n \cdot \left[\sum_{i=1}^{M} L(d(X_n) \mid \omega_i) f(X_n \mid \omega_i, B_i) Q_i \right] \right\} f(B \mid Y_{n-1}). \tag{23}$$

It is well known that for a 0, 1 loss function, the decision function minimizing (23) is

$$d(X_n) : \text{choose } \omega_j \text{ such that}$$
$$f(X_n, \omega_j \mid Y_{n-1}) = \operatorname{Sup}_i \{f(X_n, \omega_i \mid Y_{n-1})\}. \tag{24}$$

For this loss function, minimum conditional risk is equivalent to minimum conditional probability of error.

When X_n is a discrete random vector, the decision function for 0, 1 loss function is

$$d(X_n) : \text{choose } \omega_j \text{ such that}$$
$$P(X_n, \omega_j \mid Y_{n-1}) = \operatorname{Sup}_i \{P(X_n, \omega_i \mid Y_{n-1})\}. \tag{25}$$

B. Computation of $f(B \mid Y_{n-1})$ for Mixtures

In order to minimize sample-conditional probability of error, $f(B \mid Y_{n-1})$ must be computed using a priori knowledge a), b), c), and d) of Section II—C and $F(B)$—at least an appropriately defined uniform c.d.f., not ruling out the true value of B.

Working with density functions rather than c.d.f.'s, $f(B \mid Y_{n-1})$ is given by Bayes Theorem as follows:

$$f(B \mid Y_{n-1}) = \frac{f(X_{n-1} \mid B)(B \mid Y_{n-2})}{f(X_{n-1} \mid Y_{n-2})}. \tag{26}$$

The denominator on the right side of (26) is a normalization constant which assures that $f(B \mid Y_{n-1})$ integrates over the B space to unity. $f(B \mid Y_{n-2})$ is the density in the B space at the $(n-2)$ stage. $f(X_{n-1} \mid B)$ is a function directly utilizing the a priori knowledge above. It can be shown that the identifiability requirement c) is sufficient to guarantee the existence of an estimator for B_0, converging to B_0 with probability one, under rather general conditions. For example, Patrick [18] has exhibited a minimum distance estimator $\xrightarrow{P} B_0$ when $F(X_s \mid \omega_i, B_i)$ is continuous in X_s and B_i. Blischke [11] has exhibited consistent moment estimators for the parameters of a mixture of two ($M = 2$) binomial c.d.f.'s. His result immediately extends to a mixture of two ($M = 2$) multinomial c.d.f.'s.

Further research is required to determine more general conditions under which identifiability assures the existence of an estimator converging to B_0 with probability one. When such an estimator does exist, we are assured that $f(B \mid Y_{n-1})$ converges to a Dirac delta function [17] at B_0.

Using a priori knowledge a) and b) to construct $f(X_{n-1} \mid B)$ according to (9), (26) becomes

$$f(B \mid Y_{n-1}) = \frac{\left[\sum_{i=1}^{M} f(X_{n-1} \mid \omega_i, B_i)Q_i\right] f(B \mid Y_{n-2})}{f(X_{n-1} \mid Y_{n-2})}. \quad (27)$$

Equation (27) is a fundamental result for the a posteriori probability density of the vector B characterizing the parameter-conditional mixture. Note that B contains all parameters characterizing this nonsupervisory problem, including the mixing parameters $\{Q_i\}_{i=1}^{M}$.

Using the iterative solution for $f(B \mid Y_{n-1})$ in (27), the decision equation (24) is implemented by computing

$$f(X_n, \omega_i \mid Y_{n-1}) = \int Q_i f(X_n \mid \omega_i, B_i) f(B \mid Y_{n-1})\, dB \quad (28)$$

and the decision equation (25) is implemented by computing

$$P(X_n, \omega_i \mid Y_{n-1}) = \int Q_i P(X_n \mid \omega_i, B_i) f(B \mid Y_{n-1})\, dB. \quad (29)$$

C. Systems Minimizing Sample Conditional Probability of Error

In this section we consider the design of systems minimizing sample-conditional probability of error. When $F(X_s \mid \omega_i)$ is approximated by a multinomial c.d.f. using the fixed bin model, X_s is a discrete random vector. We therefore use decision equation (25) with $P(X_n, \omega_i \mid Y_{n-1})$ computed in terms of $f(B \mid Y_{n-1})$ by (29).

Using the fixed bin model, denote the bins that the v samples on the nth observation fall in by I_{η_k}, $k = 1, 2, \cdots, v$. Using this notation, we obtain from (12) and (29) that

$$P(X_n, \omega_i \mid Y_{n-1}) = v! \int \left[\left\{\prod_{k=1}^{v} p_{\eta_k}^i\right\} Q_i\right] f(B \mid Y_{n-1})\, dB. \quad (30)$$

It is convenient to define the sample-conditional expectation of

$$v!\left[\left\{\prod_{k=1}^{v} p_{\eta_k}^i\right\} Q_i\right] \text{ by } a_{n-1}^i; \text{ that is}$$
$$a_{n-1}^i = E\left[\left\{v! \prod_{k=1}^{v} p_{\eta_k}^i\right\} Q_i \mid Y_{n-1}\right]. \quad (31)$$

If $v = 1$, (31) reduces to

$$a_{n-1}^i = E[p_\eta^i Q_i \mid Y_{n-1}] \quad (32)$$

where p_η^i is the probability X_n falls in bin I_η given that ω_i caused X_n.

Equation (31) used along with decision equation (25) has an interesting interpretation: To minimize sample-conditional probability of error while making a decision on the nth sample, observe the relative frequency vector V_n. Then compute the probability of this relative frequency vector given the past samples Y_{n-1} and ω_i, for each i, and make decisions as follows:

$$\text{choose } \omega_i \quad a_{n-1}^i = \operatorname*{Sup}_i \{a_{n-1}^i\}. \quad (33)$$

When $F(X_s \mid \omega_i, B_i)$ is continuous in X_s and B_i, the decision equation is (24). For example, if the family is multinomial Gaussian,

$$f(X_{n-1} \mid \omega_i, B_i) = \frac{1}{((2\pi)^{1/2} |\Phi_{xx}^i|^{1/2})} \cdot \exp\{-\tfrac{1}{2}(X_{n-1} - \theta_i)^T [\Phi_{xx}^i]^{-1} (X_{n-1} - \theta_i)\}$$

where

$$\Phi_{xx}^i = E[X_s^T X_s \mid \omega_i] \quad \text{and} \quad \theta_i = E[X_s \mid \omega_i].$$

Note that

$$B_i = (\Phi_{xx}^i, \theta_i)$$
$$B = (\{\Phi_{xx}^i\}_{i=1}^{M}, \{\theta_i\}_{i=1}^{M}, \{Q_i\}_{i=1}^{M}).$$

The two types of optimum systems which make decisions with minimum sample-conditional probability of error are shown in Fig. 4. The upper system uses the fixed bin model, assuming the family $\{F(X_s \mid \omega_i)\}$ is multinomial and the lower system is for cases where the family is known and whose ith member is continuous in B_i and X_s.

Through (27)–(29), we have the minimum sample-conditional probability of error solution for this nonsupervisory problem and have thus obtained the second goal of the paper.

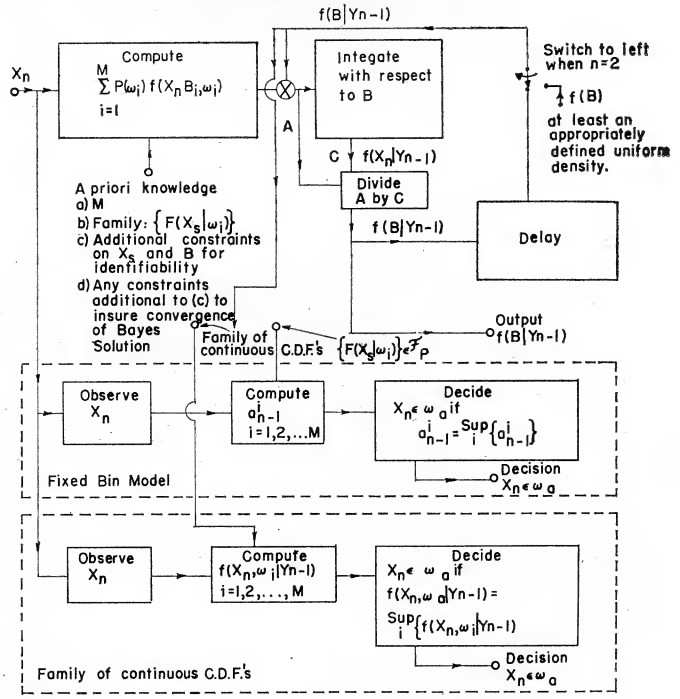

Fig. 4. Minimum conditional probability of error systems.

D. Marginal a posteriori Density of a Parameter in B

Sometimes it is desirable to obtain the a posteriori probability of just one parameter in B. The Bayes estimator of one parameter, for example, is calculated from the marginal a posteriori probability density of that parameter. Therefore, let γ_{k_i} be some parameter in B_k. The sample-conditional density of γ_{k_i} is obtained by

integrating (27) with respect to all parameters in B not equal to γ_{k_j}. Integrating (27) in this fashion gives

$$f(\gamma_{k_j} \mid Y_{n-1}) = \frac{\left[\sum_{i \neq k} \int Q_i f(X_{n-1} \mid B_i, \omega_i) f(B \mid Y_{n-1})\, d\bar{B}\right]}{f(X_{n-1} \mid Y_{n-2})} + \frac{\int Q_k f(X_{n-1} \mid B_k, \omega_k) f(B \mid Y_{n-2})\, d\bar{B}}{f(X_{n-1} \mid Y_{n-2})} \quad (34)$$

where $\bar{B}$ is defined as the vector not containing parameter γ_{k_j} but containing all other parameters in B.

Using the fact that

$$f(X_{n-1} \mid B_i, \omega_i) = f(X_{n-1} \mid B_i, \omega_i, \gamma_{k_j})$$

and

$$Q_k f(X_{n-1} \mid B_k, \omega_k) = f(X_{n-1}, \omega_k \mid \bar{B}, \gamma_{k_j}, Y_{n-2}),$$

and defining "weighting coefficients,"

$$C_i(\gamma_{k_j}) = \frac{f(X_{n-1}, \omega_i \mid Y_{n-2}, \gamma_{k_j})}{f(X_{n-1} \mid Y_{n-2})},$$

(34) becomes

$$f(\gamma_{k_j} \mid Y_{n-1}) = \left[\sum_{i \neq k} C_i(\gamma_{k_j}) + C_k(\gamma_{k_j}) \cdot \frac{E[f(X_{n-1}, \omega_k \mid \gamma_{k_j}, Y_{n-2})]}{f(X_{n-1}, \omega_k \mid \gamma_{k_j}, Y_{n-2})}\right] f(\gamma_{k_j} \mid Y_{n-2}) \quad (35)$$

where

$$E[f(X_{n-1}, \omega_k \mid \gamma_{k_j}, Y_{n-2})] = \int f(X_{n-1}, \omega_k \mid \bar{B}_k, \gamma_{k_j}) f(\bar{B} \mid \gamma_{k_j}, Y_{n-2})\, d\bar{B}.$$

The interpretation of (35) is as follows:

a) $\sum_{i \neq k} C_i(\gamma_{k_j})$ is the probability, given γ_{k_j} and Y_{n-2}, that pattern source ω_k did not cause X_{n-1}. With probability $\sum_{i \neq k} C_i(\gamma_{k_j})$, $f(\gamma_{k_j} \mid Y_{n-2})$ is retained at the $(n-1)$st stage.

b) $C_k(\gamma_{k_j})$ is the probability given γ_{k_j} and Y_{n-2}, that pattern source ω_k caused X_n. With probability C_k, $f(\gamma_{k_j} \mid Y_{n-2})$ is updated, assuming X_n is caused by pattern source ω_k.

c) $E[f(X_{n-1}, \omega_k \mid \gamma_{k_j}, Y_{n-2})]$ is involved in (35) because $f(X_{n-1}, \omega_k \mid B_k)$ is, in general, a function of parameters other than γ_{k_j}.

IV. Implementation and Computer Simulation

A. Quantizing the Parameter Space

The computation of $f(B \mid Y_n)$ is iterative, in terms of $f(B \mid Y_{n-1})$. The procedure is that, upon receiving X_n, $f(B \mid Y_{n-1})$ is replaced in storage by $f(B \mid Y_n)$. To store $f(B \mid Y_{n-1})$ in memory as mass density in the parameter space, it is necessary that B take on a finite number of points. For some cases, it is not necessary to store $F(B \mid Y_{n-1})$ this way; instead, a sufficient statistic is retained [17].

In general, however, it is necessary to store $f(B \mid Y_{n-1})$ in memory, as indicated above, if digital techniques are used. To do this, denote the number of scalar entries in B by q and write

$$B = (\phi_1, \phi_2, \cdots, \phi_q). \quad (36)$$

Quantize ϕ_i into N_i one-dimensional segments of length Δ_i each, $i = 1, 2, \cdots, q$. There are then $\prod_{i=1}^{q} N_i$ cubes, each q-dimensional. Denote a particular cube by $L_{j_1, j_2, \cdots, j_q}$, and denote the fixed but unknown probability measure in this cube by $m_{j_1, j_2, \cdots, j_q}$. Denote the a posteriori probability measure in this cube at the nth stage by $(m_{j_1, j_2, \cdots, j_q})_n$. Then, assuming $f(X_s \mid B)$ is constant as B varies within a cube, we approximate (27) by

$$(m_{j_1, j_2, \cdots, j_q})_n = \frac{f(X_{n-1} \mid L_{j_1, j_2, \cdots, j_q})(m_{j_1, j_2, \cdots, j_q})_{n-1}}{\sum_{i=1}^{q} \sum_{j_i=1}^{N_i} f(X_{n-1} \mid L_{j_1, j_2, \cdots, j_q})(m_{j_i, j_2, \cdots, j_q})_{n-1}}. \quad (37)$$

We have investigated (37) as an approximation to (27) using the IBM 7094 Computer. The example reported here is a binary ($M = 2$), one-dimensional example (see Fig. 2) where the family $\{F(X_s \mid \omega_i)\}$ is known Gaussian. A uniform random number generator was used to generate the mixing parameter Q_{1_0}. If the number from the uniform random number generator was $\leq Q_{1_0}$, mean θ_{1_0} was added to a number generated from a Gaussian random number generator with zero mean and variance σ_0. If the number the uniform random number generator was $> Q_{1_0}$, then mean θ_{2_0} was added to the number from the Gaussian generator.

For this example,

$$B_1 = (\theta_1, \sigma)$$

$$B_2 = (\theta_2, \sigma)$$

$$B = (\theta_1, \theta_2, \sigma, Q)$$

$$f(X_s \mid \omega_i, B_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{[-(X_s - \theta_i)^2/2\sigma^2]}.$$

The fixed but unknown vector B_0 is

$$B_0 = (\theta_{1_0}, \theta_{2_0}, \sigma_0, Q_{1_0}).$$

And according to Proposition 1, a sufficient constraint for identifiability is $\theta_{2_0} > \theta_{1_0}$.

First, computer simulated results were obtained for average error, using decision equation (24) and (37) for only θ_{2_0} and θ_{1_0} unknown. For $\sigma_0 = 1$, $Q_{1_0} = \frac{1}{2}$ and both known, and the constraint $\theta_{2_0} > \theta_{1_0}$ and 90 segments of length 1/10 in each dimension of the parameter space (8100 two-dimensional cubes, 5040 having zero measure because of the constraint $\theta_{2_0} > \theta_{1_0}$), the average error is plotted vs. n in Fig. 5 for the following 3 cases:

Case 1: $\theta_{1_0} = 0$, $\theta_{2_0} = 2.4$, and $F(\theta_1, \theta_2)$ uniform—i.e., the same measure is put in each of the 3960 cubes to begin the iteration.

Case 2: $\theta_{1_0} = 0$, $\theta_{2_0} = 0.5$, and $F(\theta_1, \theta_2)$ uniform.

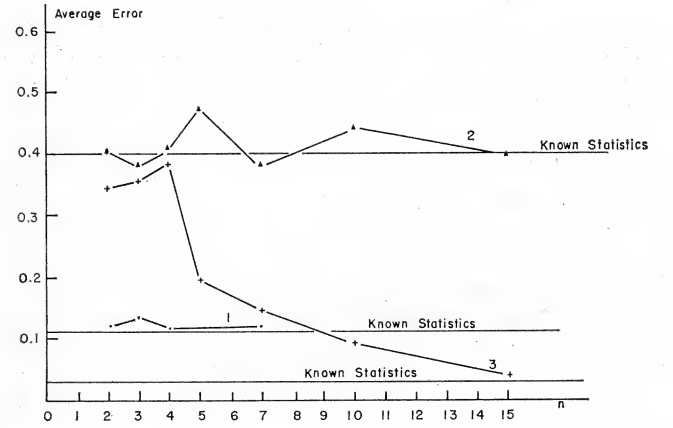

Fig. 5. Computer simulated average error vs. n for binary Gaussian example with two unknowns, θ_{1_0} and θ_{2_0}.

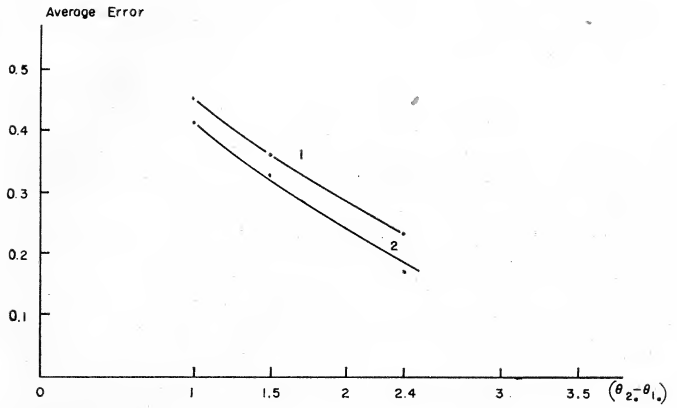

Fig. 6. Computer simulated average error vs. $(\theta_{2_0} - \theta_{1_0})$ for binary Gaussian example with three unknowns, θ_{1_0}, θ_{2_0}, and Q_{1_0}.

Case 3: $\theta_{1_0} = 0$, $\theta_{2_0} = 2$, and $f(\theta_1, \theta_2) = (1/2\pi) \exp (\theta_1 - 1/2)^2 \exp (\theta_2 - 5/2)^2$.

Case 3 has unfavorable a priori knowledge.

Second, σ_0, θ_{1_0}, and θ_{2_0} were unknown at the receiver. For $Q_{1_0} = \frac{1}{2}$ and the constraint $\theta_{2_0} > \theta_{1_0}$ and 45 segments along the θ_1 and θ_2 axis and 10 along the σ axis, all of length 1/10, and with $F(\theta_1, \theta_2, \sigma)$ uniform, average error vs. $(\theta_{2_0} - \theta_{1_0})$ is plotted in Fig. 6 for two cases:

Case 1: $n = 20$, 10 experiments, $\theta_{1_0} = 0$, $\sigma_0 = 1$, θ_{2_0} variable.

Case 2: $n = 50$, 10 experiments, $\theta_{1_0} = 0$, $\sigma_0 = 1$, θ_{2_0} variable.

There was a total of 20 250 cubes in the parameter space, with zero measure in 10 570 cubes because of the constraint $\theta_{2_0} > \theta_{1_0}$.

V. Conclusions

The approach taken in this paper to nonsupervised adaptation begins by expressing the c.d.f. of a sample X_s as a mixture c.d.f. which is given in terms of members of a family $\{F(X_s \mid \omega_i)\}$ and mixing parameters $\{Q_i\}_{i=1}^{M}$.

As a first result, for the case when $F(X_s \mid \omega_i)$ is unknown, we approximate it by a multinomial c.d.f. under the frame work of a "fixed bin" model. This can be considered an application of the histogram concept to nonsupervisory problems. The resulting construction technique provides for taking into account such nonparametric a priori knowledge as $F(X_s \mid \omega_i)$ is symmetrical and/or differs from $F(X_s \mid \omega_j)$, $j \neq i$, only by a translational vector.

As a second result, we computed the a posteriori probability $f(B \mid Y_{n-1})$ iteratively where B contains parameters characterizing each $F(X_s \mid \omega_i)$ and the mixing parameters $\{Q_i\}_{i=1}^{M}$. Systems which minimize sample-conditional probability of error were then obtained for both the case where the family $\{F(X_s \mid \omega_i)\}$ is known and the case where the fixed bin model is used. For computer implementation of both cases, the parameter space containing B was quantized.

The concept of identifiability is introduced to mean that, given $F(X_s)$, B_0 is a unique solution of $F(X_s) = F(X_s \mid B)$ where $F(X_s \mid B)$ is constructed from a priori knowledge of M and the family $\{F(X_s \mid \omega_i)\}$. By alluding to the references, it was shown that an estimator converging with probability one to B_0 can be constructed under rather general conditions when given identifiability. Then there is a guarantee that $f(B \mid Y_{n-1})$ converges with probability one to a dirac delta function at B_0; and the sample conditional probability of error converges to the probability of error obtained when B_0 is known.

Finally, a binary ($M = 2$) one-dimensional example is given where the family $\{F(X_s \mid \omega_i)\}$ is Gaussian. By quantizing the parameter space, thus approximating the Bayes solution, computer simulated results were presented for average error vs. the number of samples.

Appendix I

Mixtures and Identifiability

Following Teicher's definition [10] of identifiability for one-dimensional mixture c.d.f.'s, we give the following definition of identifiability for l-dimensional mixture c.d.f.'s. All parameters in this Appendix are fixed parameters and not to be considered variable.

Identifiability of Mixture C.D.F.'s

Let $\mathfrak{F} = \{F(X \mid \alpha) : \alpha \in R_1^k\}$ constitute a family of l-dimensional index-conditional c.d.f.'s indexed by a point α in a subset R_1^k of Euclidean k-space R^k. Then, the l-dimensional mixture c.d.f.

$$F(X) = \int_{R_1{}^k} F(X \mid \alpha)\, dG(\alpha) \tag{38}$$

is the image under the above mapping, say $\bar{\mathfrak{F}}$, of the k-dimensional c.d.f. G (where the measure μ_G induced by G assigns measure one to R_1^k).

The c.d.f. $F(X)$ is called a mixture (or G-mixture of $\mathfrak{F}$) while G is referred to as the mixing c.d.f. Let $\mathcal{G}$ denote the class of all such c.d.f.'s G, and $\mathfrak{C}$ the induced class of mixtures $F(X)$ (given a priori the family $\mathfrak{F}$). Then $\mathfrak{C}$ will be said to be *identifiable* if $\bar{\mathfrak{F}}$ is a *one-to-one map of* $\mathcal{G}$ *onto* $\mathfrak{C}$.

$F(X)$ is called a finite mixture if its mixing distribu-

tion G, or rather the corresponding measure μ_G, is discrete and doles out positive mass to only a finite number (M) of partitions in R_1^k. Let these partitions be ω_i, $i = 1, 2, \cdots, M$, and the corresponding mass or measure be $P(\omega_i) = Q_i$, $i = 1, 2, \cdots M$. Then (38) becomes

$$F(X) = \sum_{i=1}^{M} F(X \mid \omega_i)Q_i. \tag{39}$$

C.D.F.'s Characterized by Finite Number of Parameters

Let $F(X)$ be characterized by a finite size vector set of parameters B and $F(X \mid \omega_i)$ be characterized by B_i. Then

$$F(X \mid B) = \sum_{r=1}^{M} F(X \mid \omega_i, B_i)Q_i \tag{40}$$

$$B = B_1 \cup B_2 \cup \cdots \cup B_M \cup \{Q_i\}_1^M. \tag{41}$$

Then $\mathcal{G}$ is the class of all such vectors B, and $\mathcal{H}$ the induced class of parameter-conditional mixtures $F(X \mid B)$. According to the previous definition, $\mathcal{H}$ is said to be identifiable if $\bar{\mathcal{F}}$ is a one-to-one map of B onto $\mathcal{H}$.

Thus, for a given c.d.f. $F(X)$, given the family $\{F(X \mid \omega_i)\}$, M, and identifiability, there is a unique vector B_0 such that $F(X) = F(X \mid B_0)$.

Sufficient a priori Knowledge for Identifiability

We are interested in determining sufficient amounts of a priori knowledge for identifiability. The following theorem and propositions, proven in Patrick [18], provide for determining if the a priori knowledge given in a particular nonsupervisory problem is sufficient for identifiability.

Theorem 1

Let $\mathcal{F} = \{F(X \mid \omega_i,)\}$ be a family whose ith member is characterized by B_i with transform $\phi_i(v_1, \cdots, v_l \mid B_i)$ defined for $V = (v_1, \cdots, v_l) \in S_{\phi_i}$ (the domain of definition of ϕ_i), such that the mapping $A : F \to \phi$ is linear and one-to-one. Suppose that there exists a total ordering ($\leq$) of $\mathcal{F}$ such that $F_1 < F_2$ implies: i) $S_{\phi_1} \subseteq S_{\phi_2}$, ii) the existence of some $V_1 \in \bar{S}_{\phi_1}$ (V_1 being independent of ϕ_2) such that

$$\lim_{V \to V_1} \frac{\phi_2(V)}{\phi_1(V)} = 0.$$

Then the class $\mathcal{H}$ of all finite mixtures of $\mathcal{F}$ is identifiable.

A special case of Theorem 1 is

Proposition 1: The class of one-dimensional mixtures of normal c.d.f.'s—with constraint that the family be ordered lexicographically by $N_i < N_j$ if $\sigma_i < \sigma_j$ or if $\sigma_i = \sigma_j$ but $\theta_i < \theta_j$—is identifiable.

The significance of Proposition 1 is that if the family $\{F(X \mid \omega_i)\}$ is one-dimensional Gaussian, then, given $F(X \mid B)$, there is a unique solution for B_0 if the a priori knowledge includes

a) $\sigma_i > \sigma_j$, $i < j$ or

b) if k is the smallest index such that $\sigma_k = \sigma_{k+1}$, then $m_k < m_{k+1}$

c) repeat a) and b) starting with $\sigma_{k+1} < \sigma_{k+2}$, etc.

In other words, a) $\cdots$ c) is sufficient a priori knowledge to assure identifiability. It is not the necessary a priori knowledge to assure identifiability. We can view a) $\cdots$ c) as a *constraint of the domain of definition of* B. If this constraint is utilized, then a unique solution for B_0 can be found, given the sequence of samples Y_{n-1} as $n \to \infty$.

For the binary, Gaussian, on-off case, no constraints are needed, if $\sigma = \sigma_1 = \sigma_2$ and P_I are unknown:

Proposition 2: The class of mixtures of two ($M = 2$) one-dimensional normal c.d.f.'s,

$$F(X \mid \omega_1, \theta_1, \sigma), \qquad F(X \mid \omega_2, \theta_2, \sigma),$$

with σ, θ_1, and Q_1 known, is identifiable.

The following is a proposition where we have extended Proposition 1 to the multidimensional case.

Proposition 3: Let $\{F(X \mid \omega_i)\}$ be a finite family of l-dimensional normal c.d.f.'s with $B_i = (\theta_i, \Phi_{xx}^i)$ with mean vector $\theta_i = (\theta_{i_1}, \theta_{i_2}, \cdots \theta_{i_l})$ and covariance matrix $\Phi_{xx}^i = [\sigma_{jk}^i]$. If the family is ordered lexicographically so that $N_1 < N_2 < N_3 < \cdots < N_M$ if $\sigma_{11}^2, \cdots, \sigma_{kk}^k > \sigma_{kk}^{k+1}, \cdots$, or if $\sigma_{kk}^k = \sigma_{kk}^{k+1}$ but $m_{kk} < m_{k+1,k}$, then the family is identifiable.

Identifiability and the Fixed Bin Model

The family $\{F(X \mid \omega_i)\}$ defined using the Fixed Bin Model has members which are multinomial distributions. In general, mixtures of multinomial c.d.f.'s are not identifiable because they are, in general, used to approximate c.d.f.'s about which little is known a priori. We then ask what constraints must be imposed on the multinomial c.d.f.'s to insure identifiability? The following propositions give a partial answer to this question.

Let the c.d.f. of X be a mixture of binomial c.d.f.'s, $F(X \mid \omega_i)$ characterized by v and, say, p_1^i. The c.d.f. of X is a mixture c.d.f., the corresponding parameter-conditional mixture c.d.f. being

$$F(X \mid B) = \sum_{i=1}^{M} F(X \mid v, p^i; \omega_i)Q_i. \tag{42}$$

The question is, what are sufficient conditions under which p_1^i and Q_i, $i = 1, 2, \cdots M$ can be uniquely found? The following Proposition 4 by Teicher [10] gives such sufficient conditions.

Proposition 4: Let $\{F(X \mid v, p_1^i), 0 < p_1^i < 1\}$ constitute a one-parameter family of binomial distributions, v being fixed and known. A necessary and sufficient condition that the class $\bigcup_{i=1}^{M} \mathcal{H}_i$, of all finite mixtures of at most M elements of $\mathcal{F}$ be identifiable is that $v \geq 2M - 1$.

The significance of Proposition 4 is illustrated by the binary ($M = 2$) one-dimensional example of Fig. 1. In this example, p_1^1 and p_1^2 can be uniquely found if X_s consists of at least three samples from the same pattern class.

We will now give an extension of Proposition 4 to a mixture of multinomial distribution.

Proposition 5: Let $\mathfrak{F} = \{F(X_s \mid v, \{p_\xi^i\}_{\xi=1}^R)\ 0 < p_\xi^i < 1\}$ constitute a family of multinomial distributions, v being fixed. A sufficient condition that the class $\bigcup_{i=1}^M \mathfrak{C}_i$ of all finite mixtures of at most M elements of $\mathfrak{F}$ be identifiable is that $v \geq 2M - 1$.

Symbols

Symbol	*Description*
X_s	sth l-dimensional vector sample
$X_s = \{X_{sk}\}_1^v$	sequence of v samples taken at the sth observation
ω_i	ith pattern source
$F(X_s)$	cumulative distribution function (c.d.f.) of X_s
$F(X_s \mid \omega_i)$	ith source-conditional c.d.f.
$F(X_s \mid \omega_i, B_i)$	ith source, parameter-conditional c.d.f.
$F(X_s \mid B)$	parameter-conditional c.d.f. of X_s
M	number of pattern sources
$\{Q_i\}_1^M$	M pattern source probabilities— also called mixing parameters
B_i	Set of vector parameters characterizing $F(X_s \mid \omega_i)$
B_{M+1}	$B_{M+1} = \{Q_i\}_1^M$
B	collection of all vectors in $B_1, \cdots, B_M, B_{M+1}$
$F(B)$	a priori c.d.f. for B
B_{i_0}, B_0	true vectors—fixed but unknown
	Fixed Bin Model Notations
R	number of quantizing levels for each dimension of X_s
I_ξ	bin ξ, one of the l-dimensional cubes in the sample space
p_ξ^i	probability X_{sk} falls in bin I_ξ given class ω_i caused X_{sk}
G^i	$= (p_1^i, p_2^i, \cdots, p_{R^l}^i)$
$\mathfrak{F}_P$	family of multinomial c.d.f.'s
$\mathfrak{F}_{TP}$	family of multinomial c.d.f.'s differing only by translational vectors
$\mathfrak{F}_{STP}$	family of symmetrical multinomial c.d.f.'s differing only by translational vectors
V_s	vector of relative frequency in the R^l bins due to sample $X_s = \{X_{sk}\}_1^v$; $V_s = (v_{s_1}, v_{s_2}, \cdots, v_{s_{R^l}})$.

References

[1] N. Abramson and D. Braverman, "Learning to recognize patterns in a random environment," *IRE Trans. on Information Theory* (*Supplement*), vol. IT-8, pp. 58–63, September 1962.

[2] S. C. Fralick, "The synthesis of machines which learn without a teacher," Stanford University, Stanford, Calif., Tech. Rept. 61308–9, April 1964.

[3] R. F. Daly, "The adaptive binary-detection problem on the real line," Stanford Electronics Lab., Stanford, Calif., Rept. TR 2003–3, Feburary 1962.

[4] D. B. Cooper and P. W. Cooper, "Nonsupervised adaptive signal detection and pattern recognition," *Information and Control*, vol. 7, no. 3, September 1964.

[5] G. S. Sebestyen, "Pattern recognition by an adaptive process of sample set construction," *IRE Trans. on Information Theory* (*Supplement*), vol. IT-8, pp. 82–91, July 1962.

[6] "First semi-annual research summary," School of Electrical Engineering Purdue University, Lafayette, Ind., July-December 1964.

[7] E. A. Patrick, and J. C. Hancock, "The nonsupervised learning of probability spaces and recognition of patterns," *1965 IEEE Internat'l Conv. Rec.*, pt. II.

[8] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*. New York: Macmillan, 1962.

[9] H. Teicher, "On the mixture of distributions," *Ann. Math. Stat.*, vol. 31, pp. 55–73, 1960.

[10] H. Teicher, "Identifiability of finite mixtures," *Ann. Math. Stat.*, vol. 34, no. 4, December 1963.

[11] W. R. Blischke, "Moment estimators for the parameters of two binomial distributions," *Ann. Math. Stat.*, vol. 33, pp. 444–454, 1962.

[12] H. Robbins, "The empirical Bayes approach to statistical decisions problems," *Ann. Math. Stat.*, vol. 35, pp. 1–20, 1964.

[13] D. A. S. Fraser, *Nonparametric Methods in Statistics*. New York: Wiley, 1957.

[14] S. C. Fralick, "Learning to recognize patterns without a teacher," Stanford University, Stanford, Calif., Tech. Rept. 6103–10, SEL–65–011, March 1965.

[15] D. G. Keehn, "A note on learning for Gaussian properties," *IEEE Trans. on Information Theory*, vol. IT-11, pp. 126–132, January 1965.

[16] "Second semi-annual research summary," School of Electrical Engineering, Purdue University, Lafayette, Ind., January-June 1965.

[17] J. D. Spragins, "Reproducing distributions for machine learning," Stanford Electronics Lab., Stanford, Calif., Tech. Rept. 6103–7, November 1963.

[18] E. A. Patrick, "Learning probability spaces for classification and recognition of patterns with or without supervision," Ph.D. dissertation, Purdue University, Lafayette, Ind., November 1965.

[19] H. J. Scudder, "Adaptive communication receivers," *IEEE Trans. on Information Theory*, vol. IT-11, pp. 167–174, April 1965.

[20] H. J. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Trans. on Information Theory*, vol. IT-11, pp. 363–371, July 1965.

[21] C. V. Jakowatz, R. L. Shvey, and G. M. White, Adaptive wavefrom recognition," GE Research Lab., Schenectady, N. Y., Tech. Rept. 60–RL–2353 E, May 1960.

[22] T. Kailath, "Adaptive matched filters," in *Mathematical Optimization Techniques*, R. Bellman, Ed. Berkeley: University of California Press, 1963, pp. 109–140.

[23] Proakis and Drouilhet, "Performance of coherent detection systems using decision directed channel measurement," M. I. T. Lincoln Lab., Cambridge, Mass., Rept. 646–1, June 27, 1963.